CINETRANS: LEARNING TO GENERATE VIDEOS WITH CINEMATIC TRANSITIONS VIA MASKED DIFFUSION MODELS

Xiaoxue Wu[1,2] **Bingjie Gao**[2,3] **Yu Qiao**[2†] **Yaohui Wang**[2†] **Xinyuan Chen**[2†]
[1]Fudan University , [2]Shanghai Artificial Intelligence Laboratory , [3]Shanghai Jiao Tong University
xxwu24@m.fudan.edu.cn, {gaobingjie,qiaoyu,wangyaohui,chenxinyuan}@pjlab.org.cn

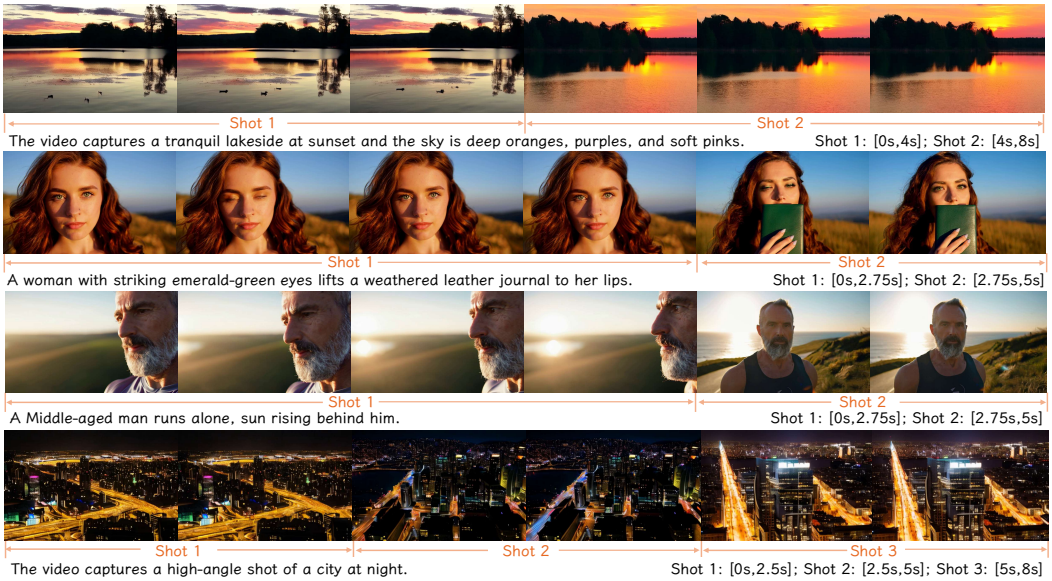


Figure 1: Multi-shot videos generated by CineTrans, which enables cinematic transitions aligning with film editing. The corresponding mask is constructed based on the timestamps of the shots, thereby controlling cinematic transitions Project page: *https://uknowsth.github.io/CineTrans/*

ABSTRACT

Despite significant advances in video synthesis, research into multi-shot video generation remains in its infancy. Even with scaled-up models and massive datasets, the shot transition capabilities remain rudimentary and unstable, largely confining generated videos to single-shot sequences. In this work, we introduce **CineTrans**, a novel framework for generating coherent multi-shot videos with cinematic, film-style transitions. To facilitate insights into the film editing style, we construct a multi-shot video-text dataset **Cine250K** with detailed shot annotations. Furthermore, our analysis of existing video diffusion models uncovers a correspondence between attention maps in the diffusion model and shot boundaries, which we leverage to design a mask-based control mechanism that enables transitions at arbitrary positions and transfers effectively in a training-free setting. After fine-tuning on our dataset with the mask mechanism, CineTrans produces cinematic multi-shot sequences while adhering to the film editing style, avoiding unstable transitions or naive concatenations. Finally, we propose specialized evaluation metrics for transition control, temporal consistency and overall quality, and demonstrate through extensive experiments that CineTrans significantly outperforms existing baselines across all criteria.

[†]Corresponding authors.

1 INTRODUCTION

Endowed with extensive pre-training and sophisticated architectures, diffusion models (Blattmann et al., 2023; Ho et al., 2020; Rombach et al., 2022; Song et al., b) have demonstrated promising capabilities in generating videos with high visual quality and strong consistency comparable to real videos (Wan et al., 2025; Kong et al., 2024; Yang et al., 2024; Li et al., 2024). Particularly in the domain of text-to-video generation (T2V) (Wang et al., 2024; Yang et al., 2024; Li et al., 2023; Cho et al., 2024), the breakthrough has attracted considerable attention, highlighting the potential of diffusion models in mastering video creation (Wang et al., 2024; Xu et al., 2024; Chen et al., 2023a; 2024a). However, generating multi-shot videos with cinematic editing style and movie-like transitions from a brief user input continues to be a significant challenge.

Existing work on long video generation can be broadly categorized into two aspects. First, larger models trained on massive datasets advance the capability of video interpretation and generation, enhancing the visual fidelity and maximum video length. (Kong et al., 2024; Wan et al., 2025; Yang et al., 2024). Second, techniques such as conditioning improve consistency across concatenated samples, enabling longer outputs through clip stitching (Zheng et al., 2024; He et al., 2023; Xie et al., 2024; Zhao et al., 2024). While both approaches partially support multi-shot sequence generation, they still exhibit notable limitations. Large-scale models primarily focus on single-shot videos due to the scarcity of shot transitions in their training datasets (Wang et al., 2023b; Ju et al., 2025; Wang et al., 2023a). Cinematic transitions are not guaranteed, let alone at precisely controlled positions. Moreover, the high computational cost and extended training time undermine the efficiency of this approach. Meanwhile, generating individual shots separately and concatenating them requires substantial manual intervention, and ignores prior knowledge from cinematic multi-shot dataset, resulting in cuts that often misalign with real-world editing styles. Additionally, many recent works often target narrow contexts, such as facial consistency (Zheng et al., 2024) or specific animated series (Dalal et al., 2025), which constrains their applicability to general video generation.

While the aforementioned methods achieve certain multi-shot effects, very few explicitly focus on cinematic transitions within diffusion models. In this work, we propose **CineTrans**, a framework that produces multi-shot videos with cinematic transitions as shown in Figure 1. To support this, we develop a refined data processing pipeline that processes raw footage into a 250K video-text dataset. Our split-stitch procedure groups semantically related clips and removes gradual transitions, and we employ transition-aware models to annotate hierarchical captions. The resulting **Cine250K** provides frame-level shot labels and temporally-dense captions, preserving film editing style and making it well-suited for multi-shot video generation, while proving effective in enabling more natural shot transitions and stronger inter-shot consistency.

To analyze how diffusion-based models handle cinematic multi-shot sequences, we dive deep into the attention patterns, examining attention maps across the temporal dimension. We find that the attention maps have strong impact on the intra-shot and inter-shot frames. This insight clarifies the underlying mechanism of shot transitions in diffusion models. Building on this, we introduce a mask mechanism featuring strong correlations within shots and weak correlations between shots, achieving controlled cinematic transitions and enabling zero-shot multi-shot generation.

As shown in Figure 2, our proposed CineTrans framework functions through two key aspects. First, through an analysis of the attention maps, we prove that the mask mechanism aligns with the diffusion model's inherent understanding of cinematic multi-shot sequences. The application of mask enables strong intra-shot frame correlations in attention module, facilitating precise frame-level cinematic transitions, which remains effective even in a training-free setting. Second, the constructed Cine250K encapsulates prior knowledge of film editing. Fine-tuning on this dataset equips CineTrans with the ability to generate cinematic transitions that conform to this style, rather than simply concatenating semantically similar clips. Consequently, CineTrans is able to producing cinematic multi-shot videos aligned with film-editing conventions.

We evaluate the model on a series of prompts with specified transitions and conduct a comprehensive analysis using multiple metrics from different perspectives. Experimental results demonstrate that CineTrans achieves finer shot transition control and stronger consistency compared to other multi-shot generation methods, without compromising overall quality.

Our contributions are summarized as follows:

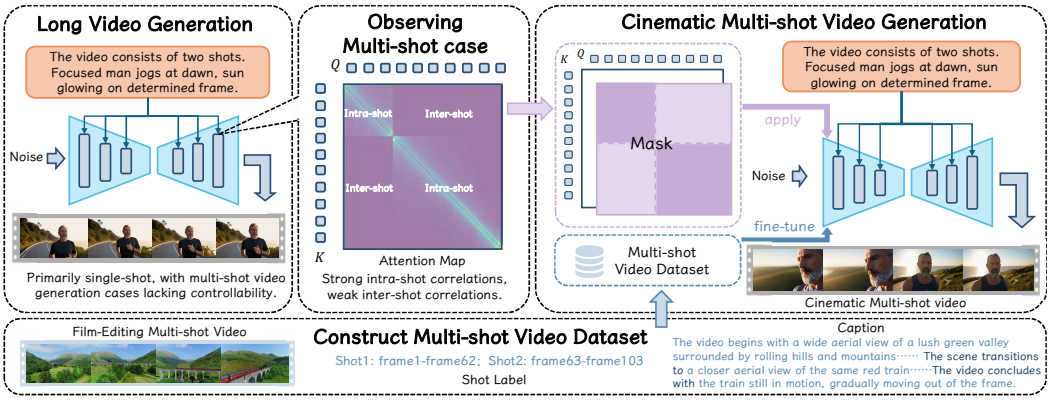


Figure 2: **Overview of CineTrans.** Existing video generation models focus primarily on single-shot video. The multi-shot video generation cases often follow several failures, remaining unstable and uncontrollable. Observations of these multi-shot cases reveal a structured pattern in attention layers. Based on this insight, we introduce a mask mechanism and fine-tune the model with our constructed dataset Cine250K, resulting in significantly improved performance.

- We developed a dataset of 250K video-text pairs, complete with frame-level shot labels and hierarchical annotations, which facilitates video diffusion models for generating cinematic transitions and consistency between shots.
- We analyze attention maps in diffusion models for multi-shot video generation and observe a strong connection between attention probabilities and shot transitions. Building on this insight, we introduce a mask mechanism that enables cinematic transitions within diffusion models, leading to the CineTrans framework, which is effective even in a training-free setting.
- We propose a series of comprehensive metrics tailored for cinematic multi-shot video generation and evaluate CineTrans, demonstrating its ability to control cinematic transitions, enhance temporal consistency, and preserve overall quality.

2 Related Work

Diffusion-based Video Generation Diffusion-based approaches, built on the iterative denoising framework of Latent Diffusion Model (Rombach et al., 2022), utilize scaled-up model (Kong et al., 2024; Wan et al., 2025; Yang et al., 2024; Blattmann et al., 2023; Chen et al., 2023a; 2024a; He et al., 2022; Lin et al., 2024; Polyak et al., 2024) and large video datasets (Miech et al., 2019; Bain et al., 2021; Zellers et al., 2021; Xue et al., 2022; Wang et al., 2023a;b; Chen et al., 2024b; Xiong et al., 2025) to generate high-quality, prompt-adherent videos with extended durations. Due to their strong semantic understanding, certain pretrained models (Kong et al., 2024; Wan et al., 2025) can preliminarily generate multi-shot videos when provided with transition-specified prompts, but at the expense of vast computational resources, long training times, and unreliable, imprecise transitions. In contrast, our work offers frame-level stable cinematic transitions, seamlessly applied to the diffusion-based framework.

Multi-Shot Video Generation. Recent work has explored multi-shot video generation, which can be categorized into two main approaches. The first generates each shot separately and then concatenates them, focusing on consistency between the generated shots. Animate-a-Story (He et al., 2023) uses motion structure retrieval for plot-aligned clips guided by text prompts. (Chen et al., 2023b) proposes a mask-based diffusion model to generate a smooth transition between shots. Zhuang et al. (2024) attempts to utilize the language model to generate prompts for different shots of a video. DreamFactory (Xie et al., 2024) employs an LLM-based framework with multi-agent collaboration and keyframe iteration design to ensure consistency and seamless transitions. MovieDreamer (Zhao et al., 2024) adopts a hierarchical autoregressive architecture for global coherence, and VGoT (Zheng et al., 2024) uses keyframes and identity-preserving embeddings to enforce temporal and character consistency. While these consistency-driven approaches are effective to some extent, they

do not leverage real multi-shot video datasets and tend to overlook complex relationships between shots, including variations in camera angles and scenes, which are crucial in human-edited videos. The second approach modifies model architectures to generate multi-shot content directly. Test-Time training (Dalal et al., 2025) adds a layer for long multi-shot generation but is tied to a specific animated series, limiting generalization. Mask^{2}DiT (Qi et al., 2025) applies shot-wise semantic masking yet focuses on text condition injection and assumes fixed shot durations. ShotAdapter (Kara et al., 2025) and LCT (Guo et al., 2025) introduce transition tokens and specialized positional encodings, respectively, yet exhibit low consistency or require large-scale training. In contrast, our method generates cinematic multi-shot videos in a single pass with flexible frame-level control, achieving strong performance even in a training-free setting and demonstrating generalizability across diverse scenarios.

3 Problem Formulation

Diffusion models (Ho et al., 2020; Song et al., a) are a class of generative models learning to reverse a diffusion process, which learn to corrupt data via a predefined noisy Gaussian process and then invert that corruption through a trained neural network. Video diffusion models simply apply this same forward-reverse framework across temporal frames, yielding coherent video sequences $\mathbf{F} = \{f_1, f_2, \ldots, f_N\}$, where each element f_t represents a video frame. A key component in these models is integrating the attention mechanism (Vaswani et al., 2017), which allows latent variables to focus on each other's relevant information and can be formalized as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are the query, key, and value matrices, and d_k is the key dimensionality.

Cinematic multi-shot video generation task also follows the framework mentioned above, aiming to generate videos with cinematic transitions that are both aesthetically pleasing and compliant with the text prompts. Additionally, with the introduction of multi-shot, the core challenges include:

- **Generation of cinematic transitions**. The generated sequence $\mathbf{F} = \{f_1, f_2, \ldots, f_N\}$ can be divided into M sub-intervals, where each $\mathbf{F}_m = \{f_{i_m}, f_{i_m+1}, \ldots, f_{i_{m+1}-1}\}$ contains frames from index i_m to $i_{m+1} - 1$, for $m = 1, 2, \ldots, M$, M is the specified shot count, and $i_1 = 1$, $i_{M+1} = N + 1$. Within sub-intervals, frames maintain visual continuity, while noticeable changes at the boundaries create cinematic transitions.
- **Consistency within and across shots**. Within a shot, maintaining visual consistency is crucial for continuity. In contrast, inter-shot consistency emphasizes high-level semantic similarity over low-level details, i.e., maintaining consistency across shots despite substantial compositional differences. This means that its evaluation is not confined to specific attributes, such as composition or facial features, but is instead guided by film-editing conventions, ensuring applicability to general scenarios.

4 Dataset

A video with cinematic transitions integrates multiple clips while preserving consistency. To capture film editing prior knowledge for multi-shot sequences, we introduce Cine250K, a dataset for cinematic video generation. As shown in Figure 3, starting from 633K richly edited videos from Vimeo[1], we design a multi-stage preprocessing pipeline to construct the annotated multi-shot dataset.

First, the transition points are identified by Pyscenedetect (Castellano, 2024), resulting in fragmented segments. Adjacent segments with high similarity, measured by ImageBind (Girdhar et al., 2023) features, are then stitched together according to predefined rules to assemble an initial collection of 16M clips containing shot transitions. We filter this collection by aesthetic score, duration, and shot count to form the preliminary video dataset. Subsequently, since shot transitions can be categorized into instantaneous hard cuts and gradual changes that blur segment boundaries, we apply TransNetV2 (Soucek & Lokoc, 2024) to detect and remove all gradual transition frames which do

[1]https://vimeo.com

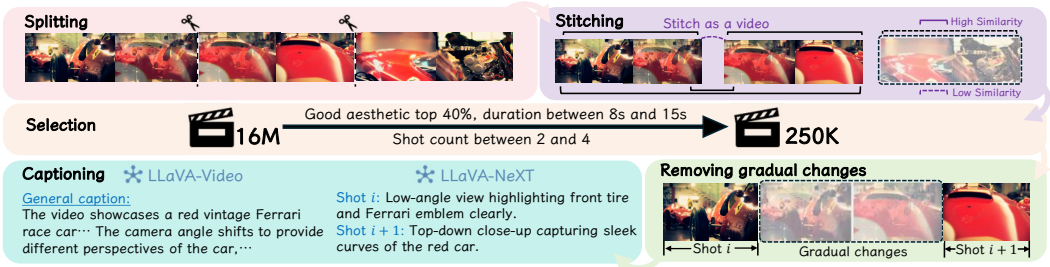


Figure 3: **Dataset curation pipeline.** The raw video is split into several clips and then selectively stitched based on semantic features. A selection process then chooses high-quality multi-shot videos. After initial assembly, gradual changes are removed. Finally, a language model is used to annotate each video with a general caption and each shot with its shot caption, yielding temporally dense annotations.

Table 1: Comparison of Cine250K and other video-text datasets.

Dataset	#Videos	Avg video len	Avg text len	Multi-shot	shot label	Aesthetic	Resolution
InternVid (Wang et al., 2023b)	234M	11.7s	17.6	✗	-	High	720P
LVD-2M (Xiong et al., 2025)	2M	20.2s	88.7	✗	-	Medium	Diverse
OpenVid-1M (Nan et al., 2024)	1M	N/A	N/A	✗	-	High	Diverse
Panda70M (Chen et al., 2024b)	70.8M	8.5s	13.2	✓	✗	Medium	720P
LLaVA-Video-178K (Zhang et al., 2024)	178K	∼ 40.4s	∼ 300	✓	✗	High	Diverse
Shot2Story20K (Han et al., 2023)	20K	16s	201.8	✓	✓	Medium	720P
Shot2Story134K	134K	N/A	N/A	✓	✓	Medium	720P
Cine250K (Ours)	250K	10.75s	148.79	✓	✓	High	720P

not clearly belong to any single shot. This yields unambiguous segments with precise shot labels, including exact start and end frame indices. In the final step, each video receives both a general caption produced by LLaVA-Video-7b-Qwen2 (Zhang et al., 2024) for temporally-dense descriptions of cinematic transitions, and separate shot captions from LLaVA-NeXT (Liu et al., 2024).

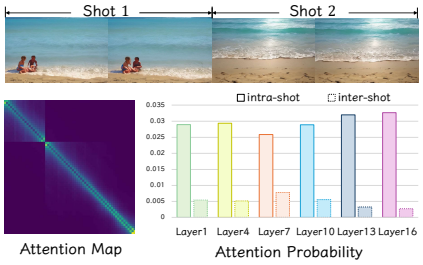
Following the pipeline, Cine250K offers high-aesthetic videos, precise shot labels, and hierarchical captions, thereby supplying rich prior knowledge for cinematic multi-shot video generation and facilitating the production of videos with authentic film editing style. In Table 1, we compare Cine250K with other datasets. As a multi-shot video dataset with detailed shot labels, the high-quality content of Cine250K can significantly facilitate research and exploration in multi-shot video generation. The details are presented in Appendix A and B.

5 Methodology

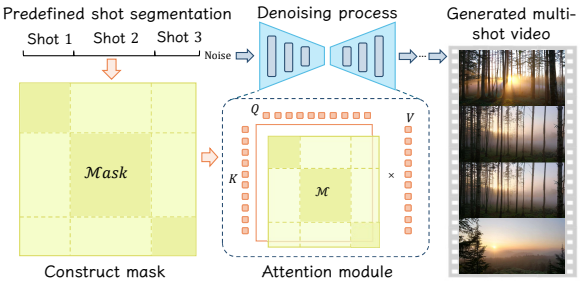
In this section, we introduce our insight and method for the proposed CineTrans. Section 5.1 presents our observations in the attention maps of diffusion models, highlighting the differences between intra-shot and inter-shot correlations. Based on this observation, we introduce a block-diagonal mask mechanism in detail in Section 5.2 for cinematic multi-shot video generation. In Section 5.3, we discuss the implementation of inference. An overview of the methodology is shown in Figure 4.

5.1 Frame Correlation in Attention Module

Recently, owing to the increased model size, dataset scale, and computational resources, some diffusion models (Kong et al., 2024; Wan et al., 2025) have demonstrated preliminary capabilities in generating multi-shot videos. However, how diffusion models internally model transitions within a video remains unclear and warrants further exploration. We hypothesize that, for correlation between two adjacent frames, there is a significant difference between transition points and non-transition points. The transition point involves a substantial shift, while a non-transition point requires continuity to maintain visual coherence, making them inherently divergent.



(a) A case of generated multi-shot video within diffusion model.



(b) The architecture of the mask mechanism.

Figure 4: We observe that in multi-shot scenarios the attention maps form a block-diagonal pattern, i.e., certain layers exhibit higher intra-shot than inter-shot frame correlations, so we design a corresponding masking mechanism. Using predefined transition points, the mask is applied to those layers of the diffusion model to guide cinematic multi-shot video generation.

In video diffusion models, the denoising process captures temporal correlations through the attention module. Specifically, the video latent representation is flattened into a sequence of tokens. Tokens from different frames participate in the calculation of the attention maps in specific layers, enabling the pretrained model to generate continuous, high-quality video segments. As a result, attention maps are a valuable tool for analyzing frame correlations.

We explore and visualize frame-wise attention maps in the context of the multi-shot video generation cases. As shown in Figure 4a, it demonstrates strong correlations for intra-shot frames and weak correlations for inter-shot frames. More specifically, the attention probability matrix exhibits a block-diagonal structure, with each block corresponding to a shot. Figure 4a also illustrates the quantitative differences in attention probabilities across various layers of the diffusion models, highlighting the variations between intra-shot and inter-shot correlations. To further quantify this observation, we compute the average ratio of mean intra-shot to inter-shot attention probabilities (26.68) and assess its correspondence with the ground-truth shot boundaries via Pearson correlation, yielding r=0.71 ($p<0.01$). This suggests the potential of leveraging the attention maps to guide the generation of multi-shot videos.

5.2 Mask Mechanism

Building on our observation, we introduce a mask mechanism, a simple strategy that operates on the attention probability in text-to-video diffusion models. Specifically, we construct an attention mask $\mathcal{M}$ for the visual tokens in the attention module at specific layers as follows:

$$\mathcal{M}_{ij} = \begin{cases} 0 & \text{if } i, j \in \text{same shot} \\ -\infty & \text{if } i, j \notin \text{same shot} \end{cases} \tag{2}$$

The mask matrix is subsequently added to the attention score in Eq. 3, effectively weakening the correlations across different shots:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathcal{M}\right)\mathbf{V}. \tag{3}$$

As a result, the final attention probabilities form block-diagonal matrices, with cinematic transitions occurring at predefined positions. As shown in Figure 4b, transition positions are specified in advance, and the mask matrix is constructed accordingly, enabling precise control over the process. With the block-diagonal mask mechanism, diffusion models can generate cinematic multi-shot videos with fine-grained control.

The mask mechanism functions through two aspects. First, it aligns with the phenomenon observed in Section 5.1, conforming to diffusion models' inherent understanding of cinematic multi-shot sequences. Second, unmasked layers enable full-frame token attention, allowing each token to attend to others across different shots, thereby establishing high-level semantic consistency. This facilitates

Table 2: **Quantitative results**. The best and runner-up are in **bold** and underlined.

Method	Transition Control Score↑	Inter-shot Consistency				Intra-shot Consistency		Aesthetic Quality↑	Semantic Consistency↑
		Semantic		Visual		Subject↑	Background↑		
		Score↑	Gap↓	Score↑	Gap↓				
StoryDiffusion +CogVideoXI2V	-	0.5214	0.4966	0.5660	0.3605	**0.9783**	0.9713	0.6296	0.2091
HunyuanVideo +Cinematron	0.3787	0.5631	0.3764	0.6053	0.2855	0.9606	0.9721	0.5978	0.2082
HunyuanVideo	0.2111	0.5723	0.4075	0.5436	0.3485	0.9476	0.9633	0.6042	0.2064
Wanx2.1-T2V-turbo	0.2355	0.6431	0.3002	0.6516	0.2333	0.9332	0.9590	0.6324	0.2046
CogVideoX	0.0324	0.5150	0.5915	0.6248	0.2226	0.9310	0.9582	0.5509	0.2061
CineTrans-Unet (Ours)	**0.8598**	**0.8095**	0.2444	0.7247	**0.1457**	0.9598	0.9725	0.5747	**0.2224**
CineTrans-DiT (Ours)	0.7003	0.7858	**0.1552**	0.7874	0.1901	0.9673	**0.9775**	**0.6508**	0.2109

Table 3: **Quantitative results** for ablation study. The best are in **bold**.

Method	Transition Control Score↑	Inter-shot Consistency				Intra-shot Consistency		Aesthetic Quality↑	Semantic Consistency↑
		Semantic		Visual		Subject↑	Background↑		
		Score↑	Gap↓	Score↑	Gap↓				
CineTrans-Unet Ablation									
w/o Mask, w/o Tuning	0	-	-	-	-	0.9615	0.9702	**0.5901**	0.2110
w/o Mask, w/ Tuning	0.2398	0.7900	0.3279	0.7226	0.2148	0.9582	0.9718	0.5711	0.2077
w/ Mask, w/o Tuning	0.6168	0.7962	0.4336	**0.8186**	0.3000	**0.9616**	0.9719	0.5764	0.2196
CineTrans-Unet	**0.8598**	**0.8095**	**0.2444**	0.7247	**0.1457**	0.9598	**0.9725**	0.5747	**0.2224**
CineTrans-DiT Ablation									
w/o Mask, w/o Tuning	0.2051	0.5924	0.3421	0.5274	0.3574	0.9153	0.9523	0.6322	0.2063
w/o Mask, w/ Tuning	0.2112	0.6532	0.3422	0.6087	0.3312	0.9213	0.9526	0.6345	0.2019
w/ Mask, w/o Tuning	0.6564	0.7838	0.1772	0.7844	0.1943	0.9618	0.9746	**0.6556**	0.2093
CineTrans-DiT	**0.7003**	**0.7858**	**0.1552**	**0.7874**	**0.1901**	**0.9673**	**0.9775**	0.6508	**0.2109**

the effective utilization of multi-shot video datasets, thereby generating multi-shot videos that align with film editing. We present the impact of applying the mask at different layers on the results in Appendix D.2.

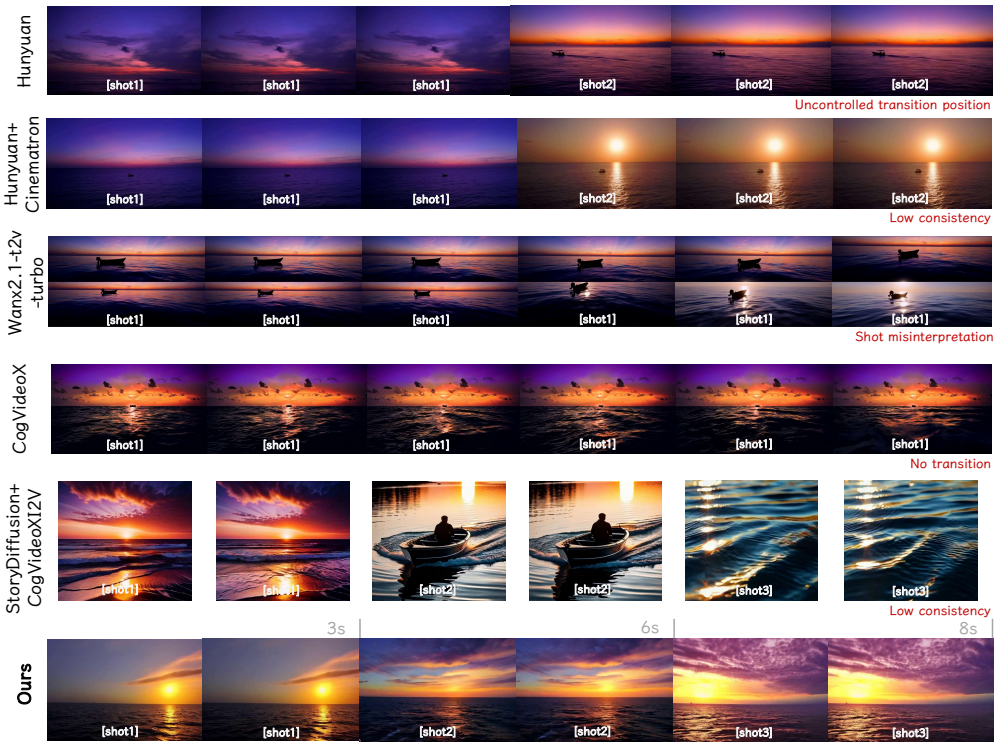
5.3 Implementations

Visible-First-Frame Attention. Beyond the block-diagonal pattern detailed in Section 5.1, we additionally observe that, in certain attention layers, all visual tokens correlate strongly with the first temporal latent. This finding suggests a pronounced reliance on initial video information. To exploit this observation, we introduce a Visible-First-Frame mechanism within the attention layers, augmenting the mask design from Section 5.2 to improve consistency in multi-shot video generation. Further details are provided in Appendix C.1.

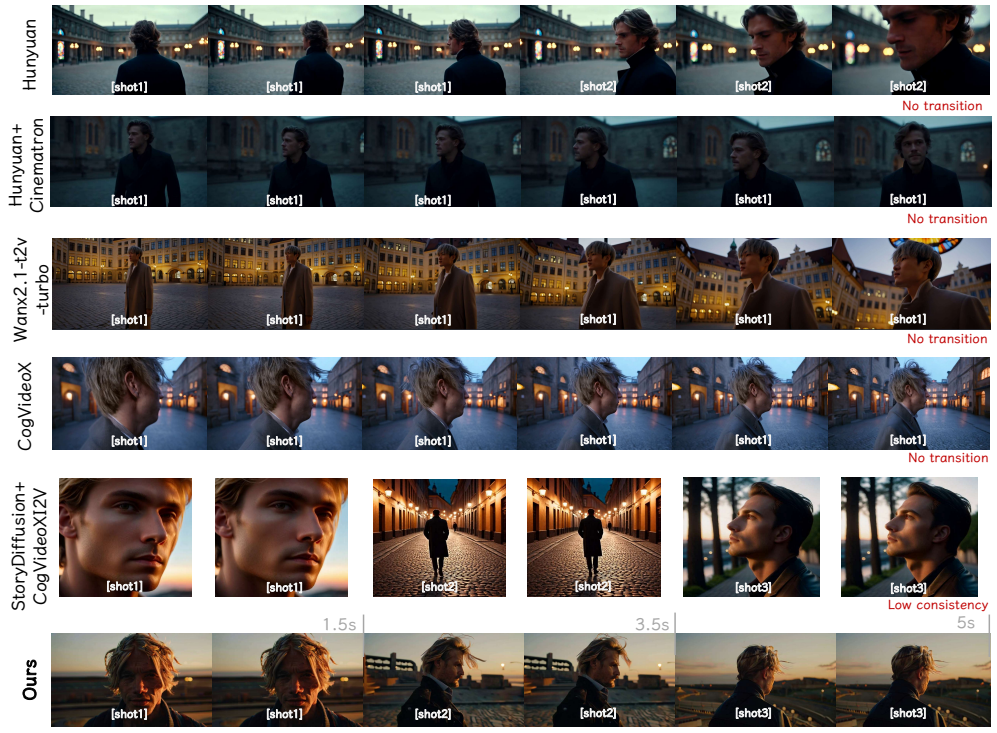
Customization. Building on our mask mechanism, we guide a diffusion model originally designed for single-shot generation to execute user-specified transitions, thus producing genuine multi-shot videos. For instance, by incorporating LoRA (Hu et al., 2022) weights, we achieve zero-shot generation of multi-shot sequences with enhanced consistency and user-defined styles or character attributes, even though those weights are originally trained on single-shot videos.

6 Experiment

In this section, we present the implementation details, evaluation settings, and results. Section 6.1 describes the details of CineTrans and baselines, which are evaluated on a series of metrics designed for the cinematic multi-shot video generation task. The metrics and results are presented in Section 6.2. Section 6.3 demonstrates that CineTrans performs well due to the components we propose.



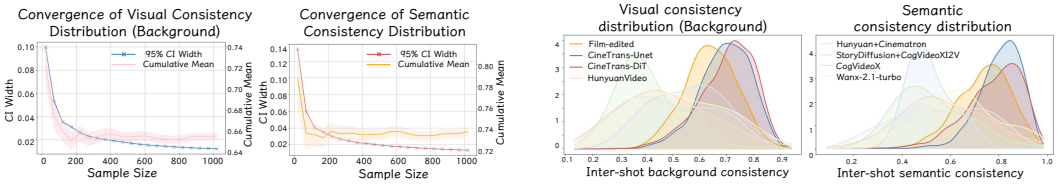
(a) A sunset ocean panorama gradually shifts to a warm close-up of sunlit clouds. (shot count: 3)



(b) A lone ash-blond traveler with angular cheekbones walks through a dusk-lit plaza. (shot count: 3)

Figure 5: **Qualitative results for different methods.** Our proposed CineTrans outperforms others in transition control while preserving coherence between shots, aligning with film-editing styles. The figure illustrates the shot segmentation results and specified shot count.

Figure 6: Training-free results of model customization using LoRA weights.



(a) Convergence of consistency distribution.

(b) Consistency distribution for the methods.

Figure 7: **Inter-shot consistency metric details and results.** (a) As the number of samples in the reference set increases, the 95% CI Width and the cumulative mean of subset converge, indicating stability in the reference set's inter-shot consistency distribution. (b) Inter-shot consistency distributions of different methods, CineTrans most closely matches that of the film-edited reference set.

6.1 Implementation Details

We implement CineTrans-Unet based on LaVie (Wang et al., 2024), and the mask mechanism is applied to the last six layers. We finetune the model on Cine250K using a total batch size of 128 for 20,000 steps. The learning rate is set to 1×10^{-4}. CineTrans-DiT extends Wan2.1-T2V-1.3B (Wan et al., 2025) by integrating a mask mechanism applied to transformer layers 7-28 and is released in two variants. The training-free variant augments the block-diagonal mask with the Visible-First-Frame Attention described in Section 5.3, enabling the generation of high-quality multi-shot videos. The second variant applies LoRA fine-tuning (rank=64) on top of the former, using a total batch size of 256 for 2,800 steps. All experiments are conducted on NVIDIA A100 GPUs. Moreover, we apply LoRA weights to CineTrans-DiT on top of the training-free version to enable model customization, and the results are shown in Figure 6.

For baseline comparison, we select three categories: large-scale T2V diffusion model, multi-shot model, and customization model. CogVideoX1.5-5B (Yang et al., 2024), HunyuanVideo Kong et al. (2024), and Wanx2.1-T2V-turbo[2] leverage large-scale pretraining and thus possess strong semantic understanding. Each of these models is prompted with a general instruction specifying the desired shot count. StoryDiffusion (Zhou et al., 2024) first produces a sequence of semantically consistent images following both the general prompt and individual shot-specific prompts, which CogVideoXI2V (Yang et al., 2024) then expands into a video. As a customization model, Cinematron[3] offers dedicated transition capabilities and employs the same sampling procedure as Hunyuan.

6.2 Evaluation

To comprehensively evaluate the multi-shot video generation task, we design 100 prompts spanning four categories using GPT-4o (Achiam et al., 2023), each incorporating cinematic transitions. For each initial general description, we manually annotate the shot count and then generate captions for each shot, thus constructing a complete prompt series. During evaluation, each method employs its required prompt format (multi-prompt or single-prompt) for sampling. To ensure a fair comparison, when using only the general description, the shot count is explicitly included in the text.

[2]https://tongyi.aliyun.com/wanxiang/

[3]https://civitai.com/api/download/models/1494601?type=Model&format=SafeTensor

Metrics. We propose a set of metrics to evaluate the generated videos from three perspectives: transition control, temporal consistency, and overall video quality. First, we evaluate transition control by performing shot segmentation (Soucek & Lokoc, 2024), comparing the detected number of shots in the generated video against the prompt's specified count, and computing the Transition Control Score as detailed in the Appendix E. Second, temporal consistency encompasses both intra-shot and inter-shot consistency. For intra-shot coherence, following VBench (Huang et al., 2024), we treat each shot as a video and compute both subject and background consistency, averaging the results across all shots. For inter-shot coherence, we evaluate from both semantic and visual perspectives. Semantically, we extract features for each shot using ViCLIP (Wang et al., 2023b) and compute cosine similarities. Visually, following VBench-Long, we compute the similarity between the middle frames of each shot to derive both subject and background similarity scores, which are then averaged to obtain the final inter-shot visual consistency. Furthermore, as discussed in Section 3, inter-shot consistency prioritizes high-level coherence aligned with film editing conventions, rather than low-level pixel similarity. Therefore, a high inter-shot consistency score, approaching the theoretical maximum, may actually indicate high pixel-level similarity, which violates the prior of multi-shot video design. To address this limitation, we introduce an auxiliary metric: Consistency Gap with Film Editing. For inter-shot consistency metrics, we collect reference score distributions from 1000 randomly selected multi-shot videos that have undergone professional film editing. The Jensen-Shannon Distance between the score distribution of generated videos and that of the reference set is computed as the Consistency Gap. Figure 7a illustrates the convergence of the consistency distribution for our selected reference set. This metric complements the raw consistency score by quantifying the deviation from natural film-editing style, thus providing a more comprehensive evaluation of video consistency. Finally, following VBench, we assess overall video quality using aesthetic quality and overall consistency, which also evaluate the visual and semantic aspects, respectively.

Results. The quantitative comparison is shown in Table 3. Even in the training-free setting, our proposed mask mechanism delivers strong transition control. In terms of consistency, CineTrans achieves high scores and closely aligns with film-editing style, which is also shown in Figure 7b. Because aesthetic quality is largely determined by the base model, CineTrans-Unet performs slightly worse in this regard, whereas CineTrans-DiT exhibits superior results. Furthermore, as shown in Table 4, the user study results demonstrate the superiority of our method in terms of user preference.

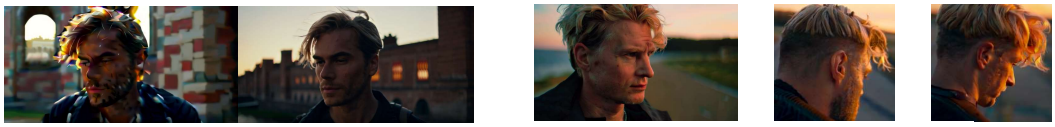
For qualitative results, as shown in Figure 5, we compare the generated results of different methods. Our method demonstrates a remarkably frame-level transition control capability while preserving coherence across different shots. Even without fine-tuning, CineTrans-DiT exhibits strong performance, demonstrating the transferability of the framework. In contrast, large-scale pretrained models fail to adhere to specified shot counts or fully misinterpret the concept of cinematic transitions. Similarly, both customized models and existing multi-shot models show low temporal consistency.

Table 4: Results of User Study

Model	Transition Control	Consistency
CineTrans-DiT (ours)	4.60 ± 0.50	$\mathbf{4.15 \pm 0.75}$
CineTrans-Unet (ours)	$\mathbf{4.75 \pm 0.44}$	4.10 ± 0.72
StoryDiffusion+CogVideoXI2V	-	3.95 ± 0.76
HunyuanVideo+Cinematron	3.60 ± 1.95	3.80 ± 0.68
HunyuanVideo	3.45 ± 1.88	3.50 ± 0.76
CogVideoX	2.50 ± 1.24	3.05 ± 0.60
Wanx2.1-T2V-turbo	3.25 ± 1.77	3.45 ± 0.69

6.3 Ablation Studies

We conduct ablation studies to assess the impact of the mask mechanism and fine-tuning process. The results shown in Table 2 demonstrate that our proposed components effectively enhance performance. In terms of inter-shot visual consistency, the fine-tuned model yields a lower Consistency Score, reflecting increased compositional variation between shots introduced by training on film-edited multi-shot videos. This effect is further corroborated by a reduced Consistency Gap, indicating closer alignment with the film-editing style. The slight decline in Aesthetic Quality after fine-tuning may be attributed to aesthetic domain differences between Cine250K and the original training set of base model. Furthermore, Figure 8 illustrates how fine-tuning and Visible-First-Frame Attention enhance video consistency and enable stable generation.



(a) Effects of Visible-First-Frame Attention. (b) Qualitative comparison of Tuning for 0.4 epoch.

Figure 8: **Qualitative comparison of key components.** (a) Applying the Visible-First-Frame Attention in CineTrans-DiT stabilizes frames. (b) Fine-tuning enhances consistency between shots.

7 Conclusion

In this paper, we introduce a novel framework CineTrans for cinematic multi-shot video generation and construct a comprehensive dataset Cine250K with detailed shot annotations. Through the analysis of attention maps in video diffusion models, we identify a strong connection between attention probabilities and cinematic transitions. Based on this observation, we propose a novel mask mechanism that enables fine-grained control over cinematic transitions, thus leading to CineTrans framework which transfers successfully in a training-free setting. Extensive experiments validate the effectiveness of CineTrans across multiple evaluation metrics, demonstrating improved transition control, temporal consistency, and overall video quality. Our work demonstrates the potential of diffusion models for multi-shot video generation, offering a new perspective for directly generating movie-like videos and paving the way for future research on controllable video synthesis.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *International Conference on Computer Vision*, pp. 1728–1738, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Brandon Castellano. Pyscenedetect: Intelligent scene-cut detection and video splitting tool (version 0.6.4). Python package and tool, 2024. URL *https://github.com/Breakthrough/PySceneDetect*. Release version 0.6.4 (June 10, 2024).

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023a.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2023b.

Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17702–17711, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025.

Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*, 2023.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *Neural Information Processing Systems*, volume 37, pp. 48955–48970, 2025.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28405–28415, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects. *arXiv preprint arXiv:2403.16407*, 2024.

Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *International Conference on Computer Vision*, pp. 2630–2640, 2019.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Maskˆ 2dit: Dual mask-based diffusion transformer for multi-scene long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18837–18846, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, b.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *ACM International Conference on Multimedia*, pp. 11218–11221, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, volume 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023a.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pp. 1–20, 2024.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023b.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788*, 2024.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. In *Neural Information Processing Systems*, volume 37, pp. 16623–16644, 2025.

Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Computer Vision and Pattern Recognition*, pp. 5036–5045, 2022.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Neural Information Processing Systems*, 34:23634–23651, 2021.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2412.02259*, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8806–8817, June 2024.

CONTENTS

Figure 9: An example of transition detection by Transnetv2. The green line represents the result of *single-frame-prediction*, while the blue line represents the result of *all-frame-prediction*.

Table 5: Threshold settings for dataset construction.

parameter		value
Pyscenedetect splitting		27
	α	0.9
	β	0.7
	γ	0.8
Transnetv2	*single-frame-threshold*	0.45
Transnetv2	*all-frame-threshold*	0.50

Table 6: Shot Segmentation Accuracy of different methods.

Method	Pyscenedetect	Transnetv2
Shot Segmentation Accuracy	65.50%	87.00%

A Details of Dataset Construction

Section 4 describes the Cine250K construction pipeline. In this section, we provide a detailed account of different stage.

A.1 Method of Splitting and Shot Segementation

During the Splitting stage, videos are initially segmented using PySceneDetect (Castellano, 2024). In the subsequent gradual-change removal stage, TransNetV2 (Soucek & Lokoc, 2024) is employed to detect and handle gradual transitions, yielding the final shot boundaries and associated shot labels. This pipeline is selected because Pyscenedetect operates exclusively on the CPU, offering greater processing efficiency than Transnetv2; however, for a pre-segmented video, Transnetv2 achieves higher shot segmentation accuracy and can handle gradual transitions more effectively than Pyscenedetect.

For Transnetv2, Figure 9 presents an example of detection. Transnetv2 provides *single-frame-prediction* and *all-frame-prediction*, which are designed to handle hard cuts and gradual transitions, respectively. The *single-frame-prediction* refers to the probability of an individual frame being a transition. In contrast, the *all-frame-prediction* predicts all frames involved in a transition, essentially estimating the probability of a frame being part of a gradual change. We apply threshold-based filtering to exclude frames with a high probability of being transition frames and obtain shot labels. Table 5 presents the threshold settings.

To quantitatively compare accuracy on shot segmentation, we randomly select 200 videos covering multiple categories, each containing multiple shots. We then apply both Pyscenedetect and Transnetv2 for shot segmentation and compare the results manually. A correctly segmented video is assigned 1; otherwise, 0. The shot segmentation accuracy is then calculated accordingly, as demonstrated in Table 6. A specific example is shown in Figure 10.

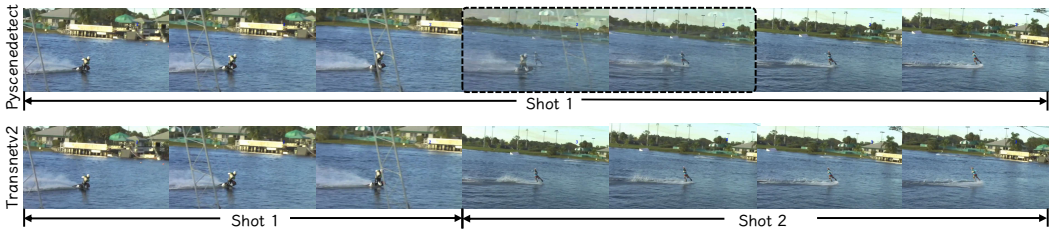


Figure 10: An example of shot segmentation by Pyscenedetect and Transnetv2. In the figure, Transnetv2, after the removing gradual changes step, accurately segments the two shots, while Pyscenedetect fails to identify the gradual changes. The transition frames caused by gradual changes are marked with dashed lines.

A.2 Details of Stitching Phase

During the Stitching stage, adjacent segments exhibiting high semantic similarity are merged to form a single video, thereby constructing multi-shot video collections. Specifically, we extract semantic features from the first and last frames of each segment using ImageBind (Girdhar et al., 2023) and quantify their similarity via Euclidean distance. For the i-th segment C^i, we denote the semantic features of its first and last frames as C^i_{first} and C^i_{end}, respectively, with their distance represented as $\text{dis}(C^i_{\text{first}}, C^i_{\text{end}})$. Based on the distance, video segments are processed sequentially and either stitched or filtered according to the following criteria.

- For a segment C^i, if $\text{dis}(C^i_{\text{first}}, C^i_{\text{end}}) > \alpha$, the segment is filtered.
- For C^i, if C^{i-1} is absent (either non-existent or filtered), C^i is treated as a video's beginning.
- For C^i, if $\text{dis}(C^{i-1}_{\text{end}}, C^i_{\text{first}}) < \beta$ and $\text{dis}(C^{i-1}_{\text{first}}, C^i_{\text{end}}) < \gamma$, then C^{i-1} and C^i are stitched to form a new segment.

Here, α restricts significant changes within a clip, β enables stitching of semantically similar transitions or originally continuous segments, and γ ensures consistency between the video's beginning and end. After processing all segments sequentially, each resulting segment is considered a complete video, forming a preliminary dataset of videos with shot transitions.

B Statistic of Cine250K

As we present in Section 4, Cine250K is a carefully curated multi-shot video dataset with detailed captions. This section presents its overall statistics. As shown in Figure 11, the average video duration is 10.75s, and the average caption length is 148.79. Most videos contain 2 to 3 shots. It is important to note that although duration and shot count filtering are applied during data processing, TransnetV2 (Soucek & Lokoc, 2024) is later utilized to remove gradual changes and re-identify shots. As a result, the final shot count and duration do not strictly adhere to the initial filtering criteria. After reidentification, 87.99% of the videos contain 2 to 5 shots. Figure 11 presents the shot distribution for videos with 1 to 10 shots, which represents 99.90% of all videos.

Regarding video categories, following Vimeo's classification, the dataset is divided into 10 categories. Among them, Travel and Documentary have relatively higher proportions. Overall, the distribution of categories is fairly balanced.

In Figure 12, we select several example video-text pairs to demonstrate the specific characteristics of the cinematic multi-shot sequences in the dataset and the style of captions.

C Additional Details of Implementation

C.1 Visible-First-Frame Attention

In Section 5.3, we introduce the implementation detail dubbed Visible-First-Frame Attention, which serves to further stabilize the mask mechanism, particularly within DiT architectures, as demon-

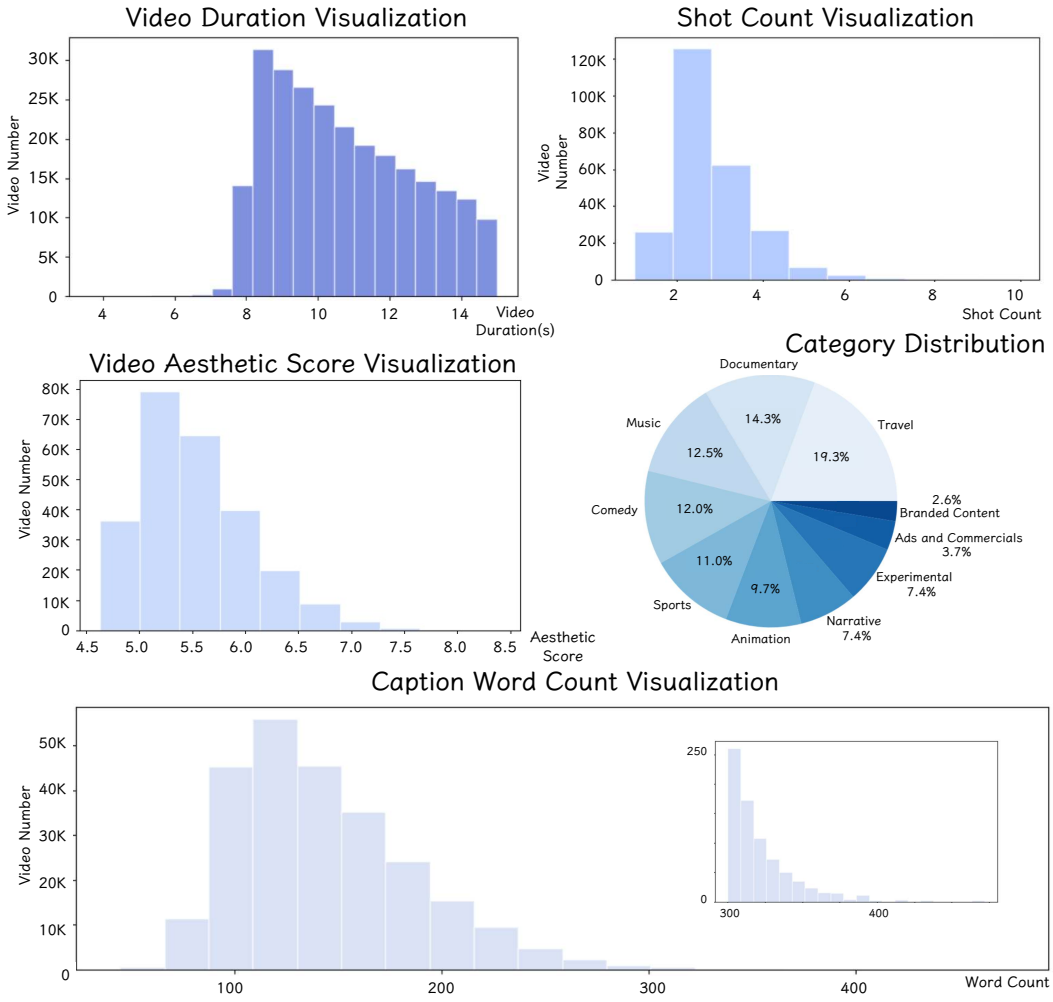


Figure 11: Statistics of Cine250K. To facilitate observation, the figure presents the shot distribution for videos containing 1 to 10 shots. The caption word count distribution only considers data with fewer than 500 words.

The video showcases a serene and picturesque setting featuring a small, red wooden cabin nestled amidst lush greenery. The cabin, with its steeply pitched roof and a chimney, is surrounded by dense foliage, including various trees and shrubs, creating a sense of seclusion and tranquility. A clear lake is visible in the background, reflecting the surrounding landscape and adding to the peaceful ambiance. The sunlight filters through the leaves, casting dappled shadows on the ground and highlighting the vibrant colors of the foliage. The camera angle shifts slightly, offering different perspectives of the cabin and its surroundings, emphasizing the natural beauty and calmness of the location.

The video features a live musical performance on stage, where a group of musicians dressed in vintage-style suits and hats are playing various instruments. The central figure, wearing a gray pinstripe suit, white shirt, red tie, and a gray fedora hat, is holding a microphone and appears to be singing or speaking. Behind him, another musician is playing a saxophone, while a third musician is playing a trombone. The stage is dimly lit with blue lighting, creating an intimate atmosphere. A banner with the text 'music is life' is visible in the background, emphasizing the theme of the performance. The musicians are deeply engaged in their performance, contributing to the overall ambiance of a live jazz or swing band concert.

The video begins with a wide aerial view of a lush green valley surrounded by rolling hills and mountains. A stone viaduct with multiple arches stretches across the valley, and a red train is seen traveling along the tracks. The train moves slowly from left to right, crossing the viaduct and moving out of the frame. The background features a clear blue sky with a few scattered clouds, and the landscape is vibrant with greenery. The scene transitions to a closer aerial view of the same red train, now clearly visible as it crosses the stone viaduct. The train continues its journey from left to right, maintaining a steady pace. The surrounding landscape remains lush and green, with rolling hills and mountains in the background under a clear blue sky. The video concludes with the train still in motion, gradually moving out of the frame.

The video features a red classic muscle car driving on a road during the late afternoon or early evening. The car, with its headlights on and license plate reading 'DL 75721', is seen from various angles as it moves forward. The background includes a mix of urban and natural elements, such as buildings, trees, and mountains under a partly cloudy sky. The setting sun casts a warm glow over the scene, enhancing the visual appeal. The car's sleek design and shiny exterior are highlighted as it navigates through the area, showcasing its smooth motion and the picturesque surroundings.

Figure 12: Example video-text pairs in Cine250K.

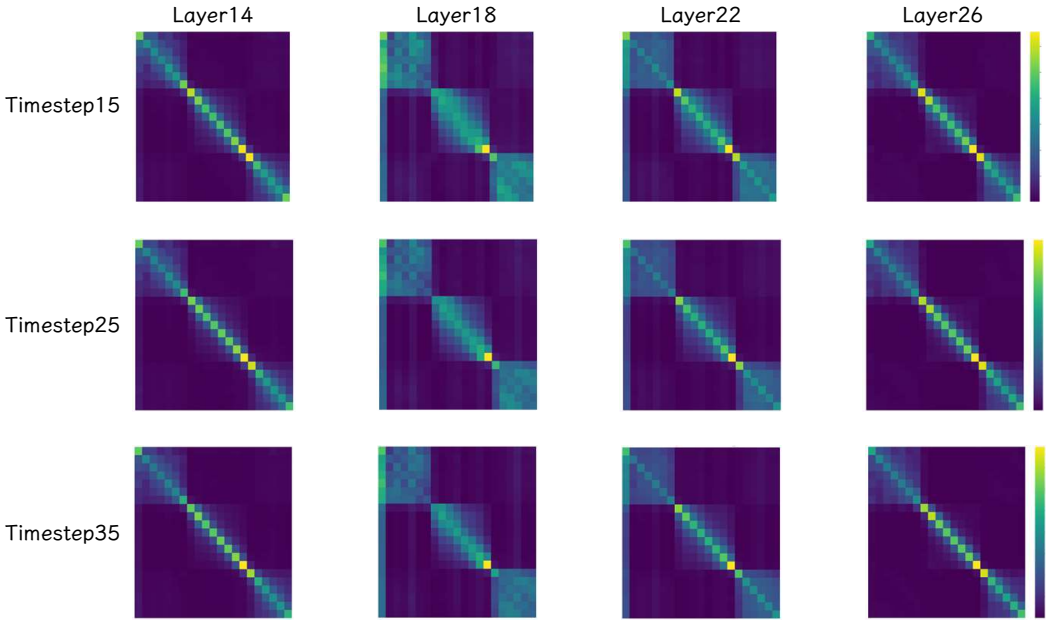


Figure 13: Visualization of the temporal-domain attention maps of Wan2.1 when generating multi-shot videos. Certain layers exhibit a pronounced focus on the first temporal slice, which motivates our Visible-First-Frame Attention mechanism.

strated in Figure 8a. This mechanism is also motivated by our analysis of DiT's attention maps during multi-shot video generation. As illustrated in Figure 13, in certain layers all visual tokens assign high attention probabilities to tokens originating from the first frame (owing to temporal compression in the VAE, this in fact corresponds to the first latent temporal slice), suggesting that the initial frame assumes a special function in the diffusion model's denoising process. In light of this observation, we modify our mask matrix so that its first column is set entirely to zero, thereby effecting the Visible-First-Frame Attention mechanism, which can be formulated as:

$$\mathcal{M}_{ij} = \begin{cases} 0 & \text{if } j = 1 \text{ or } i, j \in \text{same shot} \\ -\infty & \text{if } j \neq 1 \text{ or } i, j \notin \text{same shot} \end{cases} \tag{4}$$

C.2 Multi-prompt

In Section 6.2, we note that some methods employ a multi-prompt strategy at inference, i.e., each shot is prompted by its own text description. CineTrans-DiT also supports this capability. In addition to the primary mask matrix between video tokens, we introduce an additional mask between text and video tokens, following Qi et al. (2025), to enable precise semantic control over each shot. It is worth noting that this mask is applied only to the attention layers governing text–video token interactions, and is distinct from our proposed mask mechanism, which operates on video–video token interactions.

D Additional Results

D.1 The specific details of the attention probabilities.

In Section 5.1, we explore the frame correlations in the case of cinematic multi-shot video generation in diffusion models, where attention modules in certain layers exhibit a block-diagonal structure, i.e., strong correlations for intra-shot frames and weak correlations for inter-shot frames. In this section, we will discuss the specific details of the attention maps across different architectures, layers, and timesteps.

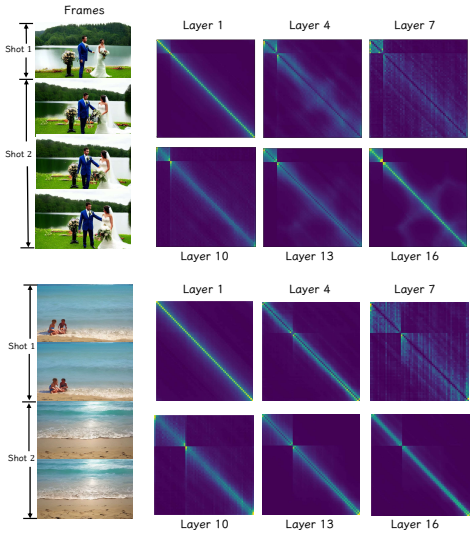
As for the architecture, we investigate both the temporal-spatial-decoupled framework and the full attention framework. The temporal-spatial-decoupled framework applies the temporal attention module directly to the time sequences, where tokens at different spatial locations do not interact with each other. In contrast, the full attention framework operates on all tokens of the video, allowing correlations across both temporal and spatial dimensions simultaneously. To focus on frame correlations, we group and average the tokens in the full attention framework by the frames before conducting further analysis. In Figure 14 and Figure 15, we use Wang et al. (2024) and Kong et al. (2024) as representatives of different architectures and present the attention maps across different layers and timesteps in both qualitative and quantitative manners.

As timesteps increase, the discrepancy between intra-shot and inter-shot attention probabilities grows in the temporal-spatial-decoupled framework, whereas it remains stable in the full attention framework. For different layers, the temporal-spatial-decoupled framework exhibits noticeable differences across most layers, whereas the full attention framework shows disparity primarily in the earlier layers. In summary, both frameworks demonstrate significant differences between intra-shot and inter-shot attention probabilities. This finding further substantiates the prevalence of strong intra-shot and weak inter-shot correlations, supporting the proposed approach.

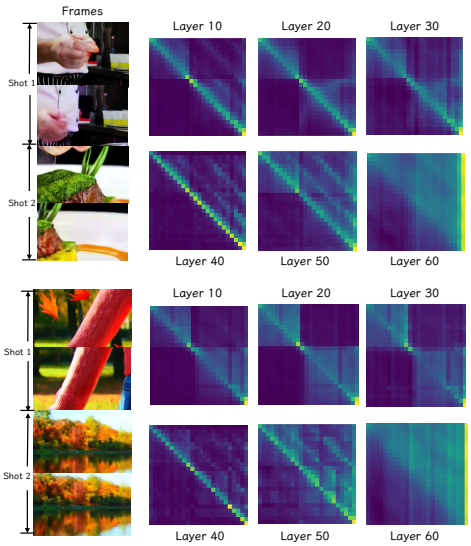
D.2 Impact of Mask Layers on Results

For CineTrans-Unet and CineTrans-DiT, the mask mechanism employs different masking layers, a design choice motivated by observations of the attention maps when diffusion models generate multi-shot videos. In this section, we examine the effects of applying masks to different layers on the quality of the generated videos, thereby demonstrating the rationale behind our masking strategy.

As shown in Figure 16a, CineTrans-Unet exhibits noticeable visual distortions when the mask is applied to all layers or only to the early layers. In severe cases, some parts of the video become indistinct or heavily degraded. When no mask is applied or when the mask is applied only to the

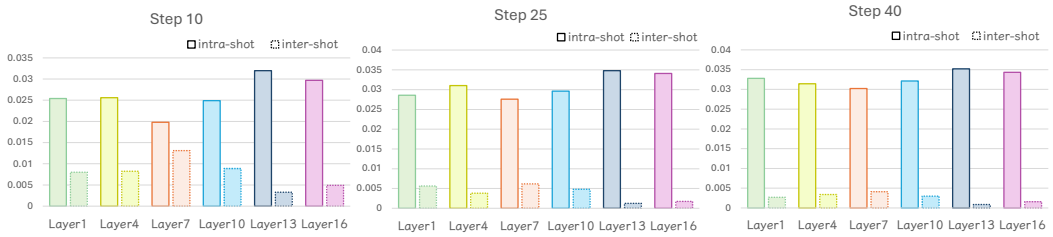


(a) Visualization of temporal attention maps in temporal-spatial-decoupled framework. The attention probability matrices across most layers exhibit a block-diagonal pattern.

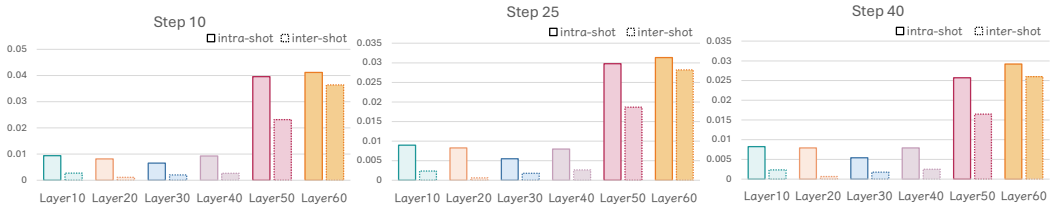


(b) Visualization of the averaged attention maps for visual tokens, grouped by frame, in the full attention framework. Earlier layers tend to exhibit a block-diagonal pattern.

Figure 14: The visualization of attention maps between visual tokens from different frames for individual case of multi-shot video generation. Both in the temporal-spatial-decoupled framework and full attention framework, diffusion models exhibit strong intra-shot attention and weak inter-shot attention in certain layers.

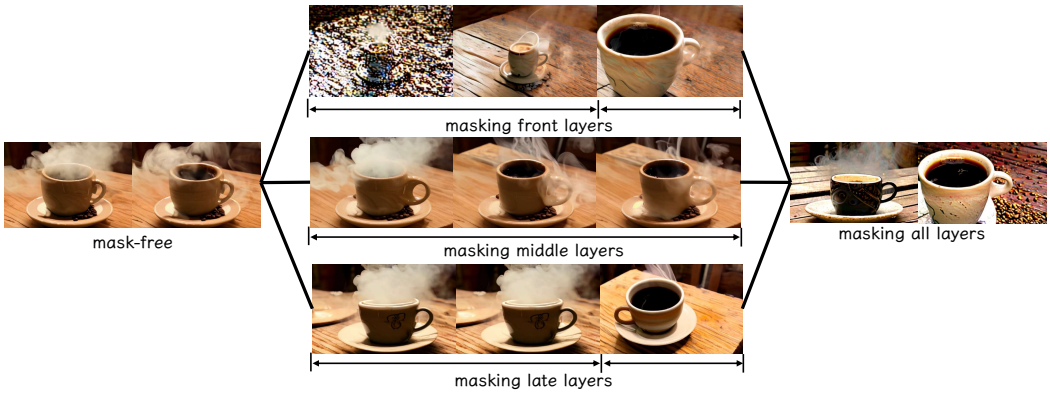


(a) Result of temporal attention probabilities in temporal-spatial-decoupled framework. In most layers, the probabilities of tokens within a shot and those across shots exhibit a noticeable difference, which seems to increase as the denoising process progresses.



(b) Result of attention probabilities for visual tokens in full attention framework. The difference between the probabilities of tokens within a shot and across shots remains relatively stable during the denoising process, with a tendency to exhibit a larger disparity in the earlier layers.

Figure 15: The average attention probability for intra-shot and inter-shot across diffusion layers and denoising steps. In both framework, the average probability within shots is obvious greater than that between shots for certain layers.



(a) Results of CineTrans-Unet.

(b) Results of CineTrans-DiT.

Figure 16: The results of different mask strategies. Applying the mask to the later layers of the spatial-temporal-decoupled architecture (CineTrans-Unet) and the middle layers of the full attention architecture (CineTrans-DiT) is considered more effective. For the full attention architecture, applying the mask to all layers leads to severe visual distortion, as illustrated in the figure.

middle layers, cinematic transitions do not emerge clearly. Applying the mask to the later layers enables effective control over transitions without significantly affecting visual quality, suggesting it as the optimal strategy. This suggests that, in the spatial-temporal-decoupled architecture, correlations in the earlier layers primarily influence visual quality, while the later layers may regulate the consistency between adjacent frames. Therefore, applying the mask mechanism to the last six layers is demonstrated to be effective.

As shown in Figure 16b, the proportion of layers requiring masking in CineTrans-DiT is larger than in CineTrans-Unet, and both the early and late layers need to retain fully visible attention. Masking all layers to this architecture may lead to low inter-shot consistency and severe visual degradation at the transitions. Conversely, masking only the earlier or later layers fails to effectively guide the transitions. Moreover, when fewer layers are masked, the inter-shot differences are reduced, which deviates from the convention of multi-shot video. These observations suggest that the current strategy of masking the middle layers achieves the optimal balance, enabling precise control over transitions while preserving a reasonable degree of consistency.

D.3 Qualitative Evaluation for Ablation Studies

The quantitative results of the ablation studies are presented in Section 6.3, and this section provides a supplementary qualitative analysis.

As shown in Figure 17, CineTrans can stably control the generation of transitions. In contrast, models without the mask mechanism, whether finetuned or not, are almost incapable of generating cinematic transitions. Direct application of the mask mechanism without further finetuning, i.e., training-free results, does indeed generate transitions. However, further training can help the model produce shot transitions that better align with the film editing style, exhibiting higher aesthetic quality and improved consistency.

E Evaluation

E.1 Prompt Details

In Section 6.2, we present a set of prompts generated by GPT-4o (Achiam et al., 2023) to evaluate the performance of multi-shot video generation. Figure 18a provides the prompt used to guide GPT.

Given the long inference time of video generation models, we design 100 prompts for evaluation, covering multiple categories. For these prompts, the transition guidance can be divided into two types. One type explicitly specifies a significant semantic change, such as prompts that include phrases like *The video transition to*, clearly indicating shot changes. The other type implicitly guides the shot transition, where the prompt does not explicitly differentiate content changes between shots, but rather provides an overall description of the video. This type of prompt design accounts for the fact that some cinematic multi-shot videos only involve switching camera angles without significant semantic change. The multi-shot guidance in such cases primarily relies on the prompt's opening phrase: *The multi-shot video consists of* $\{num\_shot\}$ *shots*. Figure 18b presents examples of these two types of prompts, and Figure 18c shows the word cloud distribution for this set of prompts. Table 7 outlines the categories of the prompts. It is worth noting that, for methods requiring multi-prompt inference, we employ GPT-4o to expand the general prompt into shot-specific captions according to the designated shot count.

Table 7: Details of prompt categories for evaluation.

category	count
scenary	34
architecture	10
human	25
object	31

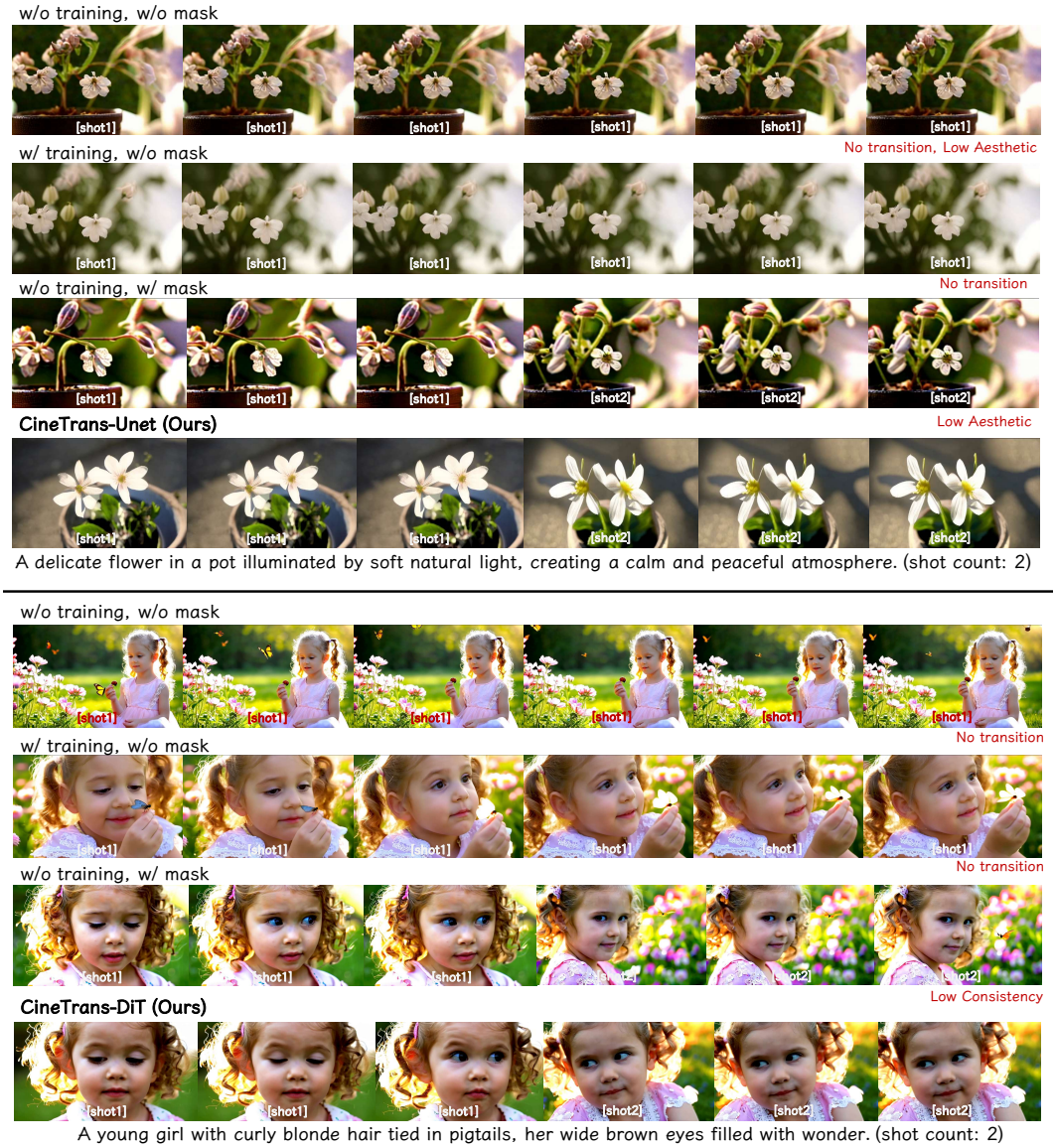


Figure 17: Qualitative results of ablation studies.

Prompt for GPT-4o

You are part of a team of bots that create videos. You are now required to generate video descriptions of content with lens transitions to be used for sampling in the video generation model. Regarding lens transitions, semantically coherent scene transformations, and changes in camera angles are allowed. The total number of words in the description is around 70 words, shot count is in the range of 2-4. The theme of the video is {*category*}.

(a)

Prompt with explicit shot change	Prompt with implicit shot change
The video begins with a comfortable living room, bathed in soft, warm light from large Windows, with plush furniture arranged around a sofa. The scene transitions to a bookshelf filled with an assortment of books, picture frames, and decorative items. The video ends with a shot of the living room as night falls, the lights dim and a peaceful atmosphere fills the space. The whole video shows a warm and comfortable scene in the living room. Shot Count: 3	The multi-shot video consists of two shots. The video captures an expansive tundra stretching beneath a pale blue sky, where the earth is hard and cracked, and sparse tufts of grass cling to life. Snow-capped mountains loom on the horizon, their peaks piercing the sky. The wind whispers across the barren landscape, carrying with it the scent of pine and cold air. In the distance, a herd of wild reindeer moves slowly across the plains, their movements graceful against the stark, frozen backdrop. Shot Count: 2

(b)

(c)

Figure 18: The details of the prompts used for evaluation. (a) Prompt for GPT-4o. The designated video theme vary. (b) Prompt examples for evaluation. (c) Word cloud of prompts for evaluation.

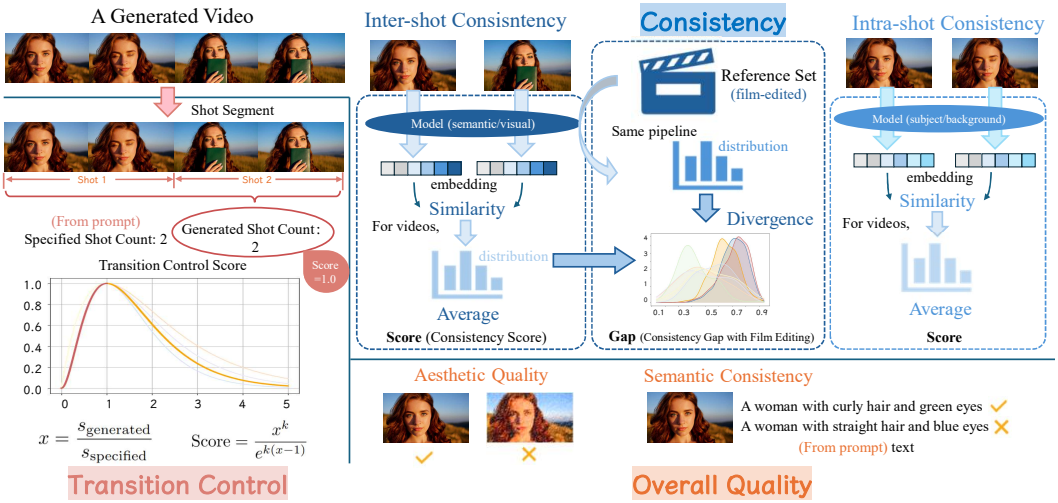


Figure 19: Overview of the metric design. We devise evaluation measures along three complementary dimensions: transition control, temporal consistency, and overall video quality.

E.2 Metric Details

In Section 6.2, we establish evaluation metrics from three aspects: transition control, temporal consistency, and overall video quality. This section will provide a detailed explanation of the specific definitions of these metrics.

Transition Control. For transition control, we define the Transition Control Score to measure whether the shot count in the generated video aligns with the specified. Specifically, the generated video is first processed by Transnetv2 for shot segmentation, yielding the number of shots, denoted as $s_{\text{generated}}$, while the prompt specifies a target shot count, $s_{\text{specified}}$. The Transition Control Score can be computed according to Equation 6.

$$x = \frac{s_{\text{generated}}}{s_{\text{specified}}} \tag{5}$$

$$\text{Transition Control Score} = \frac{x^k}{e^{k(x-1)}} \tag{6}$$

Figure 19 visualizes the calculation method of Equation 6 for the Transition Control subpanel. In practice, when $x < 1$, k is set to 2, and when $x \geq 1$, k is set to 1.6. Given that the prompts used

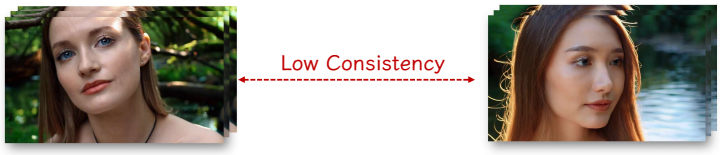


Figure 20: Failure case with low consistency, which probably results from insufficient training.

for evaluation all specify multiple transitions, the score is set to 0 when the generated video consists of a single shot. When the shot count in the generated video matches the specified value, the score is recorded as 1. More generally, the score is determined based on the absolute difference from the specified value.

Temporal consistency. In terms of temporal consistency, we consider both intra-shot consistency and inter-shot consistency. Intra-shot consistency treats each shot as a separate video and calculates the metric between adjacent frames using a method similar to that in VBench (Huang et al., 2024). The focus of this section is on inter-shot consistency. As for frame extraction, we use the middle frame of each shot for calculation. However, if the video does not generate multiple shots, inter-shot consistency cannot be evaluated. As shown in Table 2, the original LaVie (Wang et al., 2024) lacks the ability to generate transitions, and therefore its inter-shot consistency metric does not have a corresponding value.

Regarding the metric definition, inter-shot consistency cannot directly serve as the final evaluation metric. In multi-shot video generation, the goal is to ensure that the generated video aligns with the editing style of film-edited videos. High consistency would imply pixel-level similarity, which may contradict the multi-shot nature of real video editing. To address this, we extract 1000 film-edited videos as a validation dataset and compute their inter-shot consistency as a reference set. The final metric is then determined by the Jensen-Shannon Distance (JSD) between the inter-shot consistency of the generated video and that of film-edited videos, as shown in Equation 9. A lower JSD indicates a closer alignment with the reference distribution.

$$M = \frac{1}{2}(P + Q) \tag{7}$$

$$\mathrm{JS}(P \,\|\, Q) = \frac{1}{2} D(P \,\|\, M) + \frac{1}{2} D(Q \,\|\, M) \tag{8}$$

$$\mathrm{JSD} = \sqrt{\mathrm{JS}(P \| Q)} \tag{9}$$

Videos whose inter-shot consistency distribution aligns with film editing styles are considered to exhibit higher inter-shot consistency performance, while those deviating from film editing practices are regarded as having lower performance. This defines the evaluation metric based on inter-shot consistency.

F Limitation

F.1 Failure Case

The role of the mask mechanism in controlling the occurrence of transitions is relatively stable. Even in cases where the generated videos does not exactly match the expected shot count, this can be attributed to relatively small compositional differences between shots, which, despite being perceptible to the human eye as content jumps, cannot be detected by the shot segmentation model. However, CineTrans still exhibits failure cases with low consistency, as illustrated in Figure 20, which may be due to insufficient training or the need for more detailed character ID annotations in the dataset.

F.2 FUTURE WORK

Although CineTrans has achieved promising results in cinematic multi-shot video generation using the mask mechanism, there are still limitations to be addressed, along with several promising directions for future research.

- While the mask mechanism enables control over the occurrence of shots, the specification of camera viewpoints could be made more controllable, for example, enabling changes in shooting perspective within a scene that are more consistent with the film production pipeline. Therefore, achieving higher content consistency and more precise control over camera viewpoints will be a key future direction, potentially requiring the incorporation of 3D information as prior knowledge.
- Multi-shot videos could also be extended to greater lengths. Incorporating auto-regressive generation into the video generation pipeline represents a promising approach for producing longer multi-shot videos.